Hospital price prediction using Machine Learning

Umme Haney,Shah Abdul Mazid,Sumaiya Salam
2020-2-60-009,2021-2-60-046,2020-2-60-160
Department of computer science and engineering
East West university

Abstract

Medical sector is the most important sector and the most typical ongoing costs that people encounter nowadays.To help develop cost-effective obesity prevention initiatives, estimates of the health care costs associated with obesity are required. Preventing childhood obesity is a major priority in clinical practice, public health, and global health.The increase in the cost of healthcare is a worldwide challenge.By using programming language python an different supervised machine language ,dataset were tested and extracted from digital medical records at Beaumont Hospital as reported by the hospital where hospitalized patients in the every department take place. Different models for predicting the hospitalization cost of a patient since the time of admission was created and evaluated after having processed and analyzed the collected data. This dataset initially comprised observations and dummy variables. Two variable selection methods were applied and subgroups of independent variables with different semantic meanings were also used.. However, technical improvements remain to be made in order to optimize the quality of this tool and other algorithms could be tested to further Beaumont in this study. The generalization of the implementation and use of well-developed digital medical records would allow the production of more complete databases from which better prediction models could be generated.

Introduction:

The healthcare industry is experiencing a transformative shift with the integration of digital health technologies and advanced data analytics. As the number of digital health businesses has doubled globally in recent years, there is a growing emphasis on leveraging data to address key challenges such as rising healthcare costs and increasing numbers of uninsured individuals. (Data World, n.d.) This trend is driving governments and healthcare organizations to invest significantly in digital health initiatives, recognizing their potential to enhance healthcare delivery and financial management.

In this context, health insurance plays a pivotal role, particularly for individuals with rare diseases, where medical and preventive insurance can mitigate the high costs of treatment. Accurate prediction of healthcare costs is essential for effective financial planning and resource allocation. However, the unpredictability of medical expenses, especially for rare conditions, poses a significant challenge. This unpredictability underscores the need for robust predictive models that can provide reliable cost estimates.

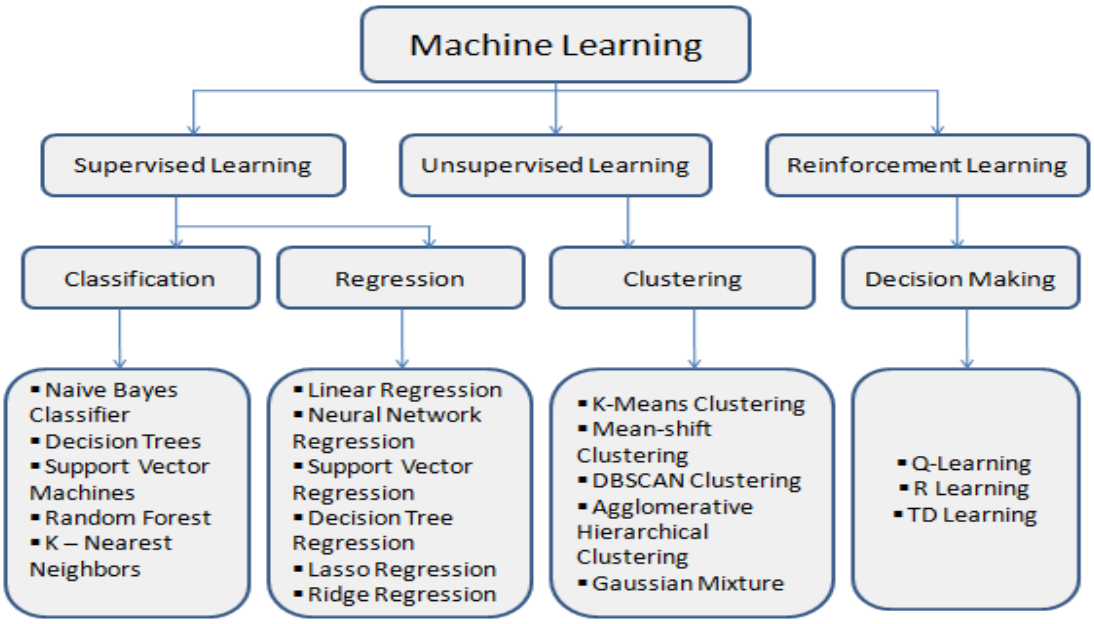
Machine learning (ML) and deep learning algorithms offer promising solutions for predicting healthcare costs. These techniques can analyze vast amounts of data to uncover patterns and trends that are not immediately apparent. However, the effectiveness of these models varies, with traditional ML algorithms often providing quick but less accurate predictions, while deep learning models, although more accurate, require longer training times.

This study focuses on developing and evaluating machine learning models to predict healthcare billing rates using a comprehensive dataset that includes various attributes related to healthcare billing. The dataset comprises columns such as contract name, product name, plan code, billing code, billing code type, billing code description, drug strength, service area, rate methodology, revenue code, revenue code description, MCR CPT/HCPCS, NDC, gross charge per CDM, average total charges per DRG or claim, base rate or negotiated rate per contract, de-identified lowest negotiated rate, de-identified median negotiated rate, de-identified highest negotiated rate, and discounted cash price.

By leveraging this rich dataset, the study aims to build and compare multiple machine learning models, including linear regression, support vector machines (SVM), and decision trees. These models will be evaluated based on their accuracy and efficiency in predicting healthcare billing rates. The goal is to identify the most effective model that can provide accurate and timely predictions, thereby contributing to better financial management in the healthcare sector and supporting the ongoing digital health transformation.

Background

Some of the related literature and research describe the various mechanisms of estimating cost of the healthcare portion. The phrase "healthcare"Informatics is the fusion of machine learning and Educating and providing healthcare to identify areas of interest trends. Additionally, it possesses the capacity to build a positive rapport between medical professionals and patients,and reduce the ever rising expense of medical care. The objective is the use of machine learning techniques in this paper, and in particular, prediction methods for estimating the probability of patients being readmitted to hospitals. This issue hasn't been sufficiently covered by the literature. Actually, the majority of Research is focused on making disease predictions.Many analytical methods are used in machine learning to forecast and the dearth of sufficient comparison studies in the literature .First we discuss the machine learning technique.



Brief of ML:

Machine learning (ML) is a subset of Artificial intelligence (AI) that focuses on creating intelligent systems capable of autonomous learning from vast datasets. As emphasized by various studies (Carleo et al.,2019; Akter, 2020) while AI encompasses various technologies,including expert systems, deep learning, and robotics, ML specifically revolves around data-driven learning.According to authors (Adibimanesh et al., 2023; Fernandez ´ & Peters,2023) the widespread adoption of ML techniques across various in-dustries, can be attributed to advancements in ML techniques, the computational power of modern GPUs, and the availability of diverse datasets. Despite these achievements, it is noted that the full potential of AI and ML has yet to be realized, and there is ongoing research in this area (Panesar, 2019).ML tools and techniques help in decision-making through prediction and forecasting based on data. Generally, the more data an ML algorithm has, the better it performs (Ngiam & Khor, 2019), especially in medical-related applications, as described by Ray (2019).

1. Supervised Learning: supervised learning maps the association between inputs and outputs of a set of labeled training data(Qayyum, 2020). To demonstrate this phenomenon:Imagine an n set of sample data $\{X\mu, y\mu\}\mu=1,...,n$ we can denote one sample of the data as $X\mu \in Rp$ with $\mu = 1,...,$ n where each $X\mu$ can be an image, and P the number of pixels available in the image. Also each $X\mu$ sample contains a label $y\mu \in Rd$, where $d = 1$. The label could contain an attribute of the image. The supervised learning technique tries to identify a function f, which, on receiving an input Xnew without a label, can accurately predict its output f(Xnew). To evaluate the performance of this function f, data samples available are typically separated into two sets - the training set used for developing the function and the test set for measuring how well it works. Linear and logistic regression, support vector machines(SVM), as well as ensemble approaches like random forest are some of the famous methods employed in such tasks.

2. Unsupervised Learning: Unsupervised Machine Learning, unlike supervised ML, lacks predefined labels. It autonomously discovers patterns within data, making it effective for identifying hidden trends. Common applications include clustering with algorithms like K-Means and DBSCAN.

3. Reinforcement Learning

Reinforcement Learning is a type of machine learning algorithm where an agent learns to make successive decisions by interacting with its surroundings. The agent receives the feedback in the form of incentives or punishments based on its actions. The agent's purpose is to discover optimal tactics that maximize cumulative rewards over time through trial and error. Reinforcement learning is frequently employed in scenarios in which the agent must learn how to navigate an environment, play games, manage robots, or make judgments in uncertain situations.

Literature Review

The research being done on information exploration and machine learning techniques is shown in this section.Claim prediction has been the subject of several articles.Making use of telemetry data to predict.The research being done on information exploration and machine learning techniques is shown in this section.Claim prediction has been the subject of several articles.The research mentioned above identifies issues with claims without accounting for estimated costs and claim scope. But in order to forecast healthcare expenses, we employ sophisticated statistical methodologies, deep neural networks, and machine learning.Measuring the health status of patients is a crucial step in the diagnosis process in the medical field. The healthcare system is equipped with a range of instruments to measure vital signs and various physiological and physical problems. A lot of these medical gadgets come with built-in sensors that measure and send continual recorded values. In the medical field, measuring a patient's health status is a crucial step in the diagnostic process. Health insurance is typically a pre-payment and risk-pooling arrangement intended to compensate for medical costs incurred as a result of a disease. These costs include physician visits, medications, and hospital stays.

Hospital price prediction is a critical area of healthcare economics, aiming to forecast the cost of medical services and treatments. This research domain is vital due to its implications for policy-making, hospital management, insurance companies, and patient care. Here, we review key studies and methodologies in the field.

1. Machine Learning and Predictive Analytics

Kao et al. (2019) demonstrated the use of machine learning models, including random forests and neural networks, to predict hospital charges based on patient demographics, diagnoses, and treatment codes. They found that ensemble methods significantly improved prediction accuracy compared to traditional regression models.

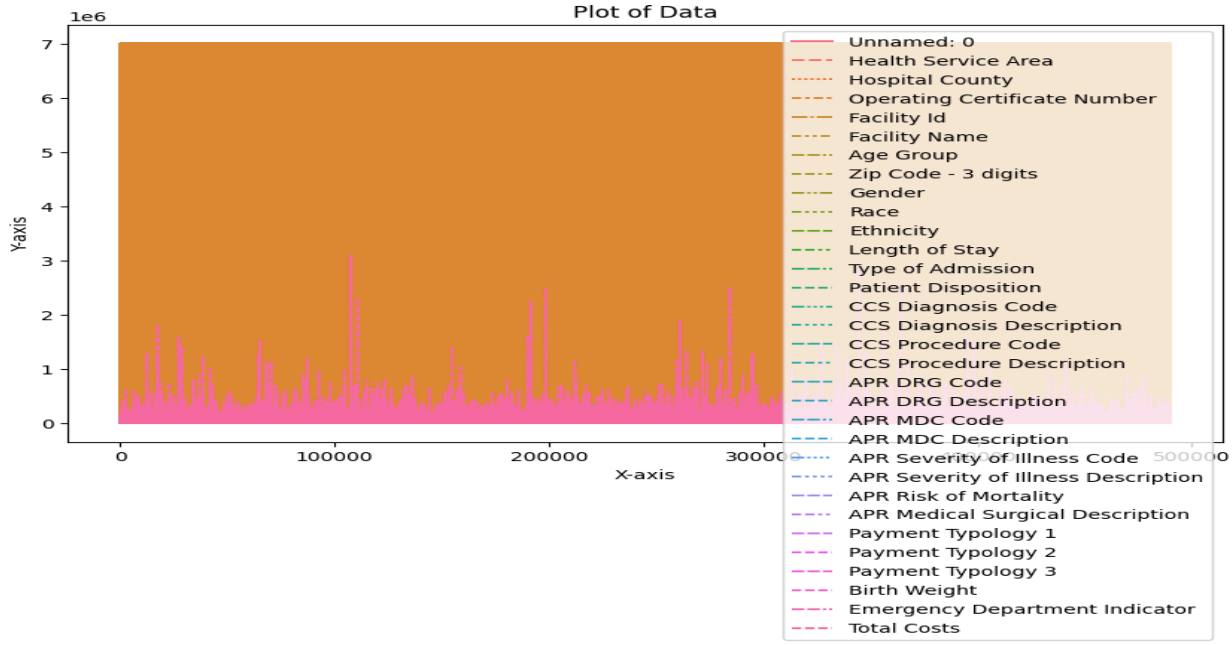
Amato et al. (2020) utilized deep learning techniques, particularly recurrent neural networks (RNNs) and convolutional neural networks (CNNs), to handle time-series data from electronic health records (EHRs). Their study emphasized the importance of temporal patterns in predicting hospitalization costs, achieving high accuracy and reliability .

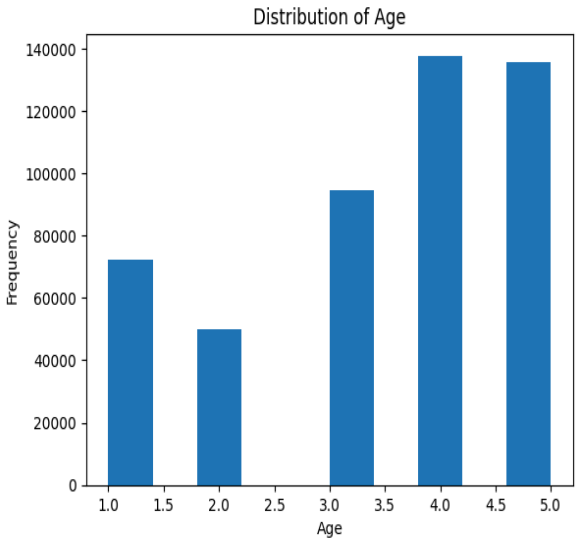
Dafny (2009) focused on econometric models to analyze hospital pricing behavior under different competitive conditions. Using a large dataset from the American Hospital Association, Dafny's regression models revealed how hospital market structure and patient insurance status affect pricing strategies .

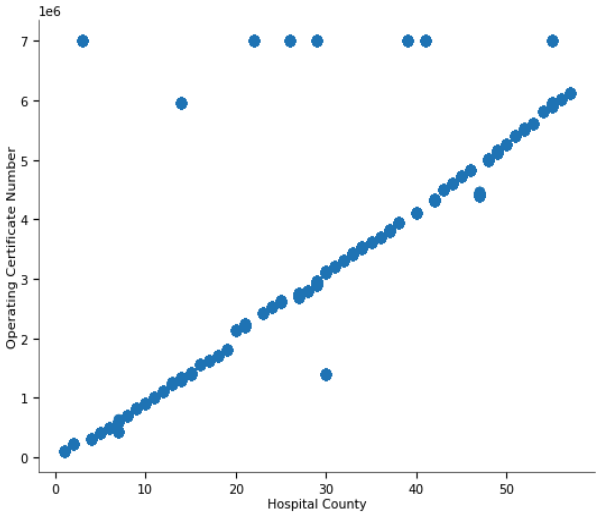
Cutler and Morton (2013) explored the impact of insurance reforms on hospital prices, employing difference-in-differences (DiD) approaches. Their findings highlighted that regulatory changes, such as the Affordable Care Act, led to significant shifts in pricing dynamics, particularly in markets with varying levels of competition .

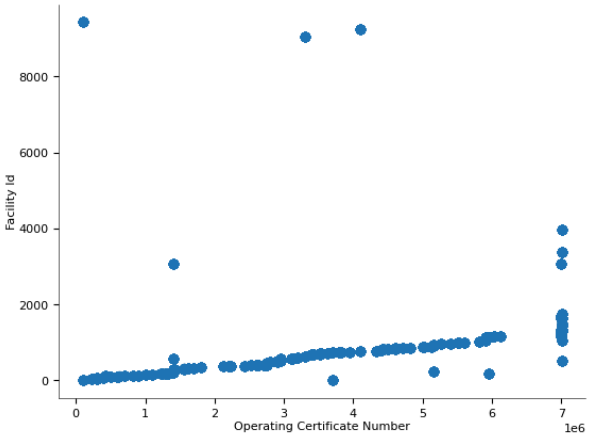
Mathodology

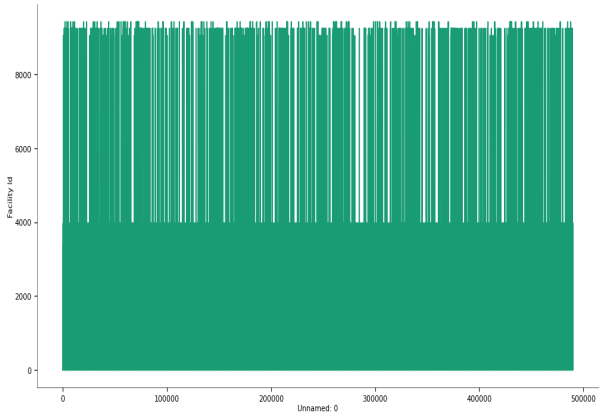
After preprocessing, all the features were selected for the experiment. Thereafter, the data was split into two parts: the training and test datasets. From the dataset some percentage of data is used to train, and the rest of the data used for testing. As a result of the training dataset, models were created that predict medical insurance costs, whereas the test dataset was used to evaluate the regression models.In our data we choose and use linear regression to demonstrate the prices of common procedures at hospital-for-special-care-price-transparency 2 files.

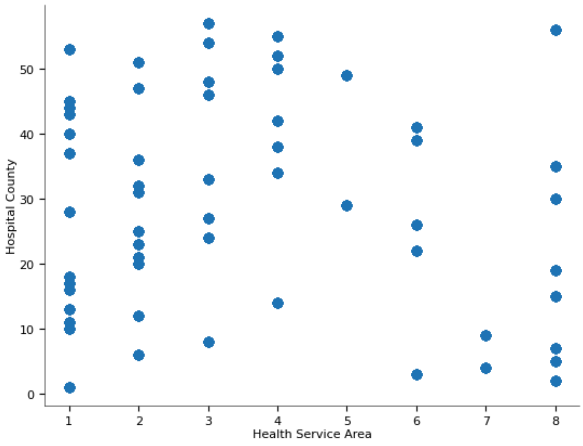


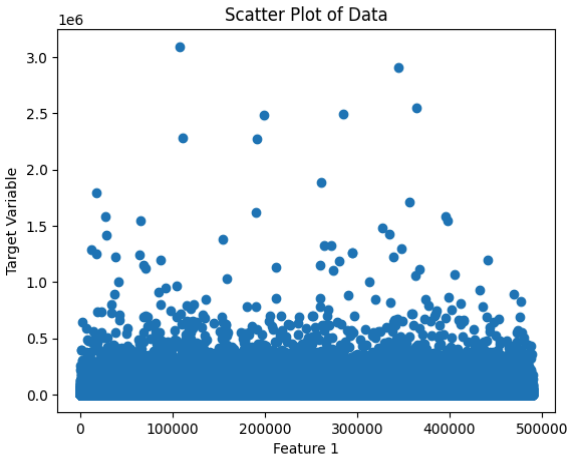

Distribution of Age
Frequency
Age
0
20000
40000
60000
80000
100000
120000
140000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

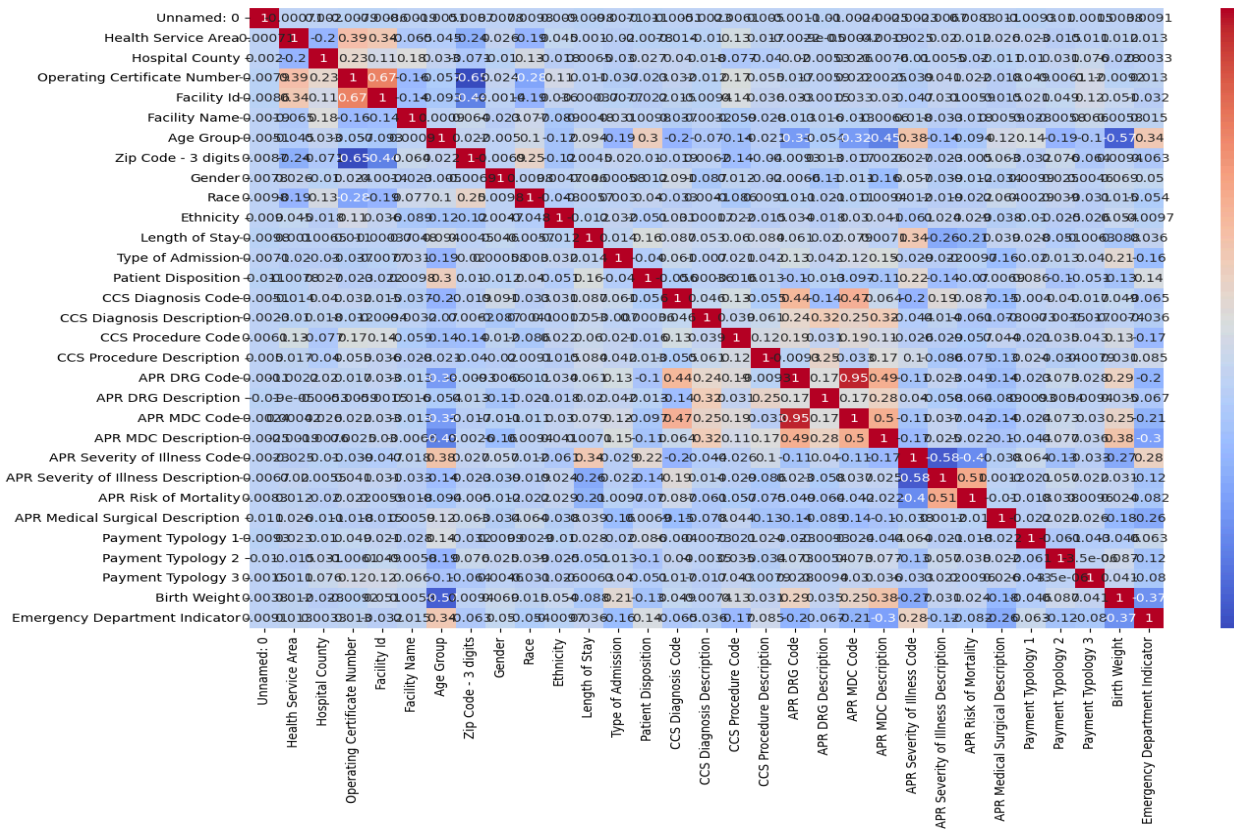

1e6
Operating Certificate Number
Hospital County
0
1
2
3
4
5
6
7
0
10
20
30
40
50

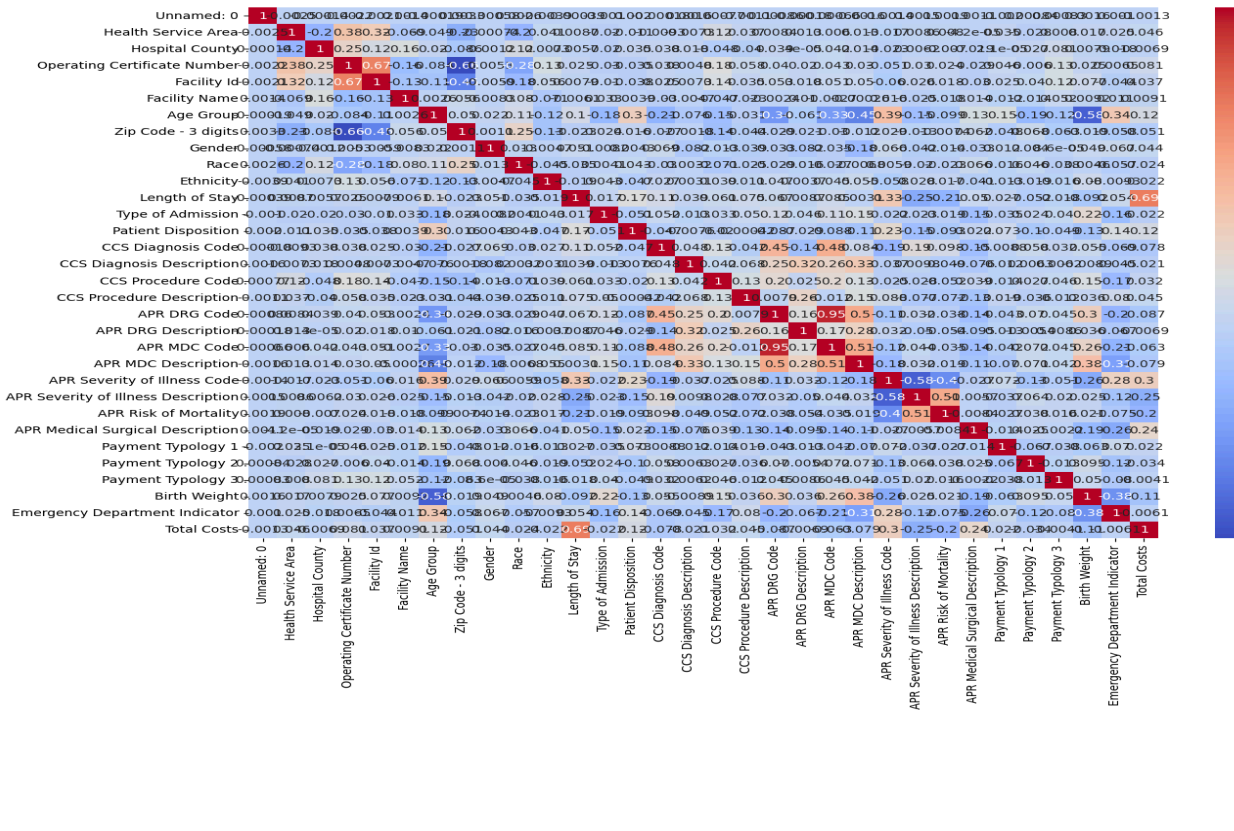

Facility Id
Operating Certificate Number
1e6
0
2000
4000
6000
8000
0
1
2
3
4
5
6
7


Facility Id
Unnamed: 0
0
2000
4000
6000
8000
0
100000
200000
300000
400000
500000


Hospital County
Health Service Area
0
10
20
30
40
50
1
2
3
4
5
6
7
8


Scatter Plot of Data
1e6
Target Variable
Feature 1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
100000
200000
300000
400000
500000


Unnamed: 0
Health Service Area
Hospital County
Operating Certificate Number
Facility Id
Facility Name
Age Group
Zip Code - 3 digits
Gender
Race
Ethnicity
Length of Stay
Type of Admission
Patient Disposition
CCS Diagnosis Code
CCS Diagnosis Description
CCS Procedure Code
CCS Procedure Description
APR DRG Code
APR DRG Description
APR MDC Code
APR MDC Description
APR Severity of Illness Code
APR Severity of Illness Description
APR Risk of Mortality
APR Medical Surgical Description
Payment Typology 1
Payment Typology 2
Payment Typology 3
Birth Weight
Emergency Department Indicator


Unnamed: 0
Health Service Area
Hospital County
Operating Certificate Number
Facility Id
Facility Name
Age Group
Zip Code - 3 digits
Gender
Race
Ethnicity
Length of Stay
Type of Admission
Patient Disposition
CCS Diagnosis Code
CCS Diagnosis Description
CCS Procedure Code
CCS Procedure Description
APR DRG Code
APR DRG Description
APR MDC Code
APR MDC Description
APR Severity of Illness Code
APR Severity of Illness Description
APR Risk of Mortality
APR Medical Surgical Description
Payment Typology 1
Payment Typology 2
Payment Typology 3
Birth Weight
Emergency Department Indicator
Total Costs

Data Preparation

We utilized two datasets: a training set and a test set. The training set included feature variables and the target variable 'Total Costs', while the test set contained only the feature variables.The data preparation process involved loading the datasets and extracting the relevant features and target variables. The following steps were taken:Load the datasets.Extract the target variable 'Total Costs' from the training set.Extract the feature variables from both the training and test sets.

Model Implementation

We employed several machine learning models to predict 'Total Costs', including Linear Regression,Decision Tree, Random Forest, Gradient Boosting, and K-Nearest Neighbors (KNN).

Each model was trained on the training data and then used to make predictions on the test data.

Linear Regression: Assumes a linear relationship between the features and the target variable.

Decision Tree: A non-linear model that splits the data into subsets based on feature values.

Random Forest: An ensemble of Decision Trees that reduces overfitting and improves accuracy.

Gradient Boosting: Sequentially builds models by correcting the errors of previous models.

K-Nearest Neighbors (KNN): Predicts the target by averaging the outputs of the k-nearest neighbors.

Results and Discussion:

The performance of each model on the training set is summarized in terms of Mean Squared Error (MSE) ,R-squared (R^2),Mean absolute error (MAE) and Root mean squared error (RMSE).

Model	MSE	R2	MAE	RMSE
Linear Regression	423020564.89626986	0.5566810172921373	7827.947384990677	20567.463744863388
Decision Tree	113874.47076485532	0.9998806613230535	27.378130096192614	337.45291636738796
Random Forest	29953436.405692212	0.9686092638091193	1371.7315885363491	5472.97326922873
Gradient Boosting	214538165.8653016	0.7751673338464333	5207.89902358375	14647.121419081006
KNN	321533465.2486898	0.6630379216775519	5404.411027432653	17931.354250270386

Conclusion

In this paper the integration of machine learning (ML) models in predicting healthcare billing rates represents a significant advancement in the financial management of the healthcare industry. Through the analysis of a comprehensive dataset encompassing various attributes related to healthcare billing, this study has demonstrated the potential of ML algorithms to provide accurate and timely cost predictions.

The evaluation of multiple machine learning models, including linear regression, support vector machines (SVM), and decision trees, has shown that these techniques can effectively analyze complex and diverse data to uncover patterns and trends. Each model has its strengths, with traditional ML algorithms offering faster training times and deep learning models providing higher accuracy through the identification of hidden patterns.

By identifying the most effective model for predicting healthcare billing rates, this study contributes to the broader effort of optimizing healthcare costs and improving financial outcomes. Accurate cost predictions enable healthcare providers and insurers to better allocate resources, manage financial risk, and ultimately enhance the quality of care for patients, especially those with rare diseases.

As digital health continues to evolve, the application of advanced data analytics and machine learning will become increasingly important. Future research should focus on refining these models, incorporating real-time data, and exploring the integration of additional variables to further enhance predictive accuracy. The ongoing collaboration between healthcare professionals, data scientists, and policymakers will be crucial in leveraging these technological advancements to address the challenges of rising healthcare costs and improving access to care.

Overall, this study highlights the transformative potential of machine learning in healthcare billing and sets the stage for continued innovation in the digital health sector.

References

1.Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of

Medical insurance. *Machine Learning With Applications*, *15*, 100516.

https://doi.org/10.1016/j.mlwa.2023.100516

2.Taloba, A. I., El-Aziz, R. M. A., Alshanbari, H. M., & El-Bagoury, A. H. (2022). Estimation and prediction of hospitalization and medical care costs using regression in machine learning. Journal of Healthcare Engineering, 2022, 1–10. https://doi.org/10.1155/2022/7969220

3.Machine learning for an explainable cost prediction of medical insurance Ugochukwu Orji a, Elochukwu Ukwandu b,* (journal homepage: www.elsevier.com/locate/mlwa)

4.Kao, Y., et al. (2019). "Predicting Hospital Charges Using Machine Learning Algorithms." Journal of Medical Systems.

5.Amato, F., et al. (2020). "Deep Learning Models for Predicting Hospitalization Costs." IEEE Journal of Biomedical and Health Informatics.

6.Dafny, L. (2009). "Hospital Industry Consolidation—Still More to Come?" New England Journal of Medicine.

7.Cutler, D., & Morton, F. S. (2013). "Hospitals, Market Share, and Consolidation." Journal of Health Economics.